Problem Statement

Introduction to GenAI and Simple LLM Inference on CPU and fine-tuning of LLM Model to create a Custom Chatbot.

Category: Artificial Intelligence, Machine Learning, LLM, NLP

Participants: 1st-4th Semester Students.

Unique Idea Brief (Solution)

The project involves creating a custom chatbot by fine-tuning the LLama 2 model using the Platypus dataset. Initially, data will be extracted from Hugging Face, preprocessed to suit LLama 2's requirements, and then uploaded back to Hugging Face. Next, the model will be fine-tuned using this dataset, aiming to optimize its performance for specific chatbot functionalities. Once fine-tuning is complete, the updated model will be uploaded to Hugging Face. Finally, a Streamlit application will be developed to provide users with a seamless interface to interact with the fine-tuned chatbot model deployed on Hugging Face.

Features Offered

- **Natural Language Understanding (NLU)**: The chatbot can understand natural language inputs from users, including questions, commands, and statements.

- **Contextual Responses**: It provides responses that are contextually relevant to the user queries, leveraging the fine-tuned LLama 2 model to generate accurate and coherent replies.

- **Multi-turn Dialogue Handling**: The chatbot can maintain context across multiple interactions, remembering previous user inputs to provide more personalized and coherent responses over extended conversations.

Features Offered

- **Intent Recognition**: It can recognize the intent behind user messages, allowing it to route queries to the appropriate functions or responses within its knowledge base.

- **Customizable Responses**: The chatbot's responses can be customized based on specific requirements or preferences, adapting to different use cases or domains.

Process flow

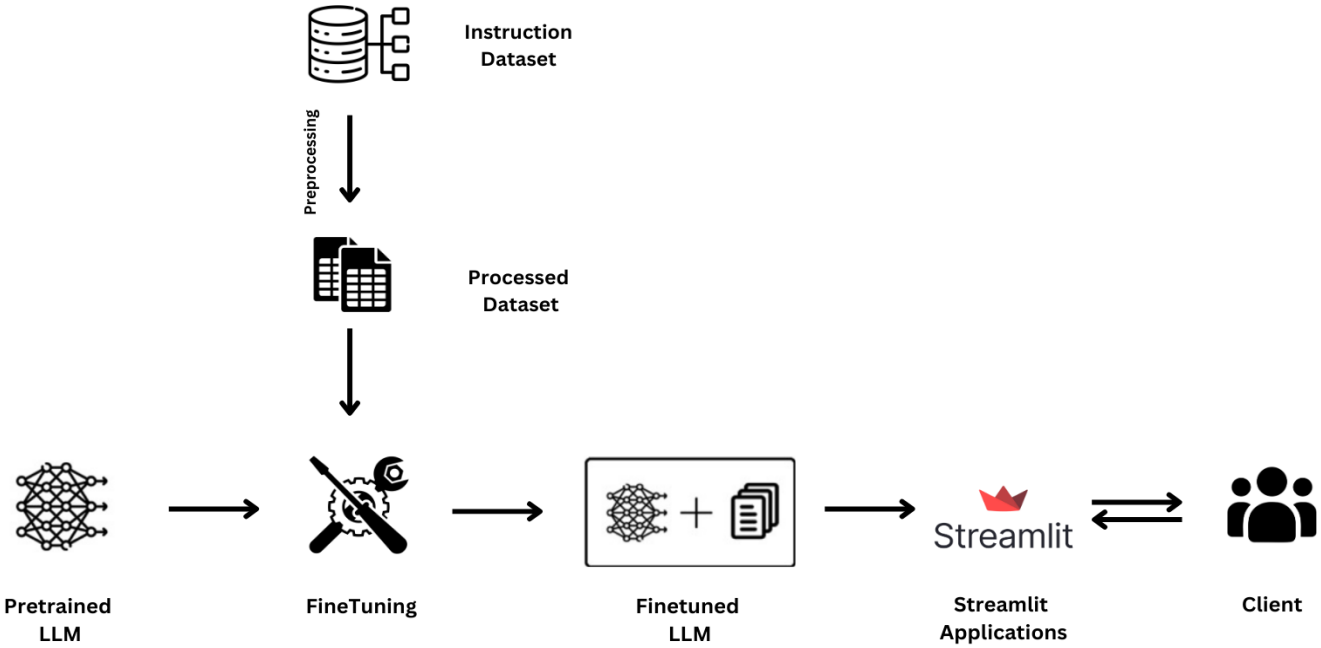
We are going to create a chatbot by fine-tuning the LLama 2 model using the Platypus dataset. The project is composed of three main steps:

i. **Extracting and Preprocessing Data:** First, we will extract data from Hugging Face and preprocess it to convert it into a suitable format for fine-tuning LLama 2. We will then upload the processed data to Hugging Face.
ii. **Fine-Tuning the Model:** In the second step, we will fine-tune the LLama 2 model using the dataset uploaded to Hugging Face. The goal is to enhance the model's performance for our specific application. After fine-tuning, we will upload the updated model to Hugging Face.

Process flow

iii. **Creating a Streamlit Application:** Finally, we will develop a Streamlit application for the chat interface. This application will allow users to interact with the fine-tuned LLama 2 model deployed on Hugging Face, providing a user-friendly chat experience.

Architecture Diagram



Technologies used

- **LLama 2 Model**
- **Python -Pytorch , Seaborn ,PEFT ..etc..**
- **Hugging Face Transformers**
- **Streamlit**
- **Hugging Face Datasets**
- **GitHub**
- **Docker**

Team members and contribution:

TEAM MEMBER 1 : THARANYA S [1024] :

Data Extraction, Preprocessing, and Upload

Extracting and Preprocessing Data:

- **Data Extraction:** Extracted data from Hugging Face datasets.
- **Data Preprocessing:** Preprocessed the data to convert it into a suitable format for fine-tuning the LLama 2 model. This involved cleaning the data, tokenization, and formatting according to the model requirements.
- **Data Upload:** Uploaded the processed data to Hugging Face.

Team members and contribution:

TEAM MEMBER 2 : THARUNIGA M [947]:

Model Fine-Tuning and Application Development

Fine-Tuning the Model and Creating a Streamlit Application:

- **Model Fine-Tuning:** Fine-tuned the LLama 2 model using the dataset uploaded to Hugging Face including setting up the fine-tuning environment, configuring hyperparameters, and running the fine-tuning process.
- **Model Upload:** Uploaded the updated model to Hugging Face after fine-tuning.
- **Streamlit Application Development:** Developed a Streamlit application for the chat interface. This involved designing the UI, integrating the fine-tuned LLama 2 model deployed on Hugging Face.

Conclusion

The initial phase involved extracting and preprocessing data from Hugging Face, ensuring the dataset was clean, tokenized, and formatted to suit the LLama 2 model's requirements. This meticulous preparation laid a strong foundation for the subsequent fine-tuning process.

The next phase saw the successful fine-tuning of the LLama 2 model using the prepared dataset. This included setting up the fine-tuning environment, configuring hyperparameters, and running the training process. The result was a model optimized for our specific application, which was then uploaded to Hugging Face for deployment.

Finally, the development of a Streamlit application provided a user-friendly chat interface, allowing seamless interaction with the fine-tuned LLama 2 model.

This interface ensures that users can easily access and benefit from the enhanced

Conclusion

capabilities of the model.

Overall, the project has achieved its goals, delivering a powerful and refined model paired with an accessible interface. This comprehensive solution is now ready to enhance user experiences and meet our specific application needs.